Electronic Letters on Computer Vision and Image Analysis 13(2):67-68, 2014

An Adaptive and Integrated Multimodal Sensing And Processing Framework For Long-Range Moving Object Detection And Classification

Tao Wang

Computer Science Department, Ph.D.
The City College of New York, 160 Convent Ave. New York, NY, USA
Advisor: Prof. Zhigang Zhu
Date and location of PhD thesis defense: 11 Dec 2012, City University of New York

Received 6 February 2014; accepted 4 Juny 2014

1 Abstract

In applications such as surveillance, inspection and traffic monitoring, long-range detection and classification of targets (vehicles, humans, etc) is a highly desired feature for a sensing system. A single modality will no longer provide the required performance due to the challenges in detection and classification with low resolutions, noisy sensor signals, and various environmental factors due to large sensing distances. Multimodal sensing and processing, on the other hand, can provide complementary information from heterogeneous sensor modalities, such as audio, visual and range sensors. However, there is a lack of effective sensing mechanisms and systematic approaches for sensing and processing using multimodalities. In this thesis, a systematical framework is proposed for Adaptive and Integrated Multimodal Sensing and Processing (AIM-SP) that integrates novel multimodal long-range sensors, adaptive feature selection and learning-based object detection and classification for achieving the goal of adaptive and integrated multimodal sensing and processing. Based on the AIM-SP framework, we have made three unique contributions. First, we have designed a novel multimodal sensor system called Vision-Aided Automated Vibrometry (VAAV), consists of a laser Doppler vibrometer (LDV) and a pair of pan-tilt-zoom (PTZ) cameras, and the system is capable of automatically obtaining visual, range and acoustic signatures for moving object detection at a large distance. It provides a close loop adaptive sensing that allows determination of good surface points and quickly focusing the laser beam of the LDV based on the target detection, surface selection, and distance measurements by the PTZ pair and acoustic signal feedbacks of the LDV. Second, multimodal data of vehicles on both local roads and highways, acquired from multiple sensing sources, are integrated and represented in a Multimodal Temporal Panorama (MTP) for easy alignment and fast labelling of the multimodal data: visual, audio and range. Accuracy of target detection can be improved using multimodalities, and a visual reconstruction method is developed to remove occlusions, motion blurs and perspective distortions of moving vehicles so that scale- and perspective-invariant visual vehicle features are obtained. The concept of MTP is not limited to visual and audio information, but is also applicable when other modalities are available that can be presented in the same time axis. With various types of features

Correspondence to: <flyingwave001@hotmail.com>

Recommended for acceptance by <Alicia Fornés and Volkmar Frinken>
ELCVIA ISSN: 1577-5097
Published by Computer Vision Center / Universitat Autonoma de Barcelona, Barcelona, Spain

extracted on aligned multimodal samples, we made our third contribution on feature modality selection using two approaches. The first approach uses multi-branch sequential-based feature searching (MBSF) and the second one uses boosting-based feature learning (BBFL). In our implementations, three types of visual features are used: aspect ratio and size (ARS), histograms of oriented gradients (HOGs), shape profile (SP), representing simple global scale features, statistical features, and global structure features, respectively. The audio features include short time energy (STE), spectral features (SPECs) which consists of spectral energy, entropy, flux and centroid, and perceptual features (PERCs) are Mel-frequency cepstral coefficients (MFFCs) for the perceptual features. The effectiveness of multimodal feature selection is thoroughly studied through empirical studies. The performance between MBSF and BBFL is compared based on our own dataset, which contains over 3000 samples of mainly four types of moving vehicles: sedans, pickup-trucks, vans and buses under various conditions. From this dataset, a subset of 667 samples of multimodal vehicle data is made publicly available at [1]. A number of important observations on the strengths and weakness of those features and their combinations are made as well.

References

[1] Audio-Visual Vehicle (AVV) dataset. http://www.vcipl.okstate.edu/otcbvs/bench/